University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2011

sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland

Dürscheid, Christa ; Stark, Elisabeth

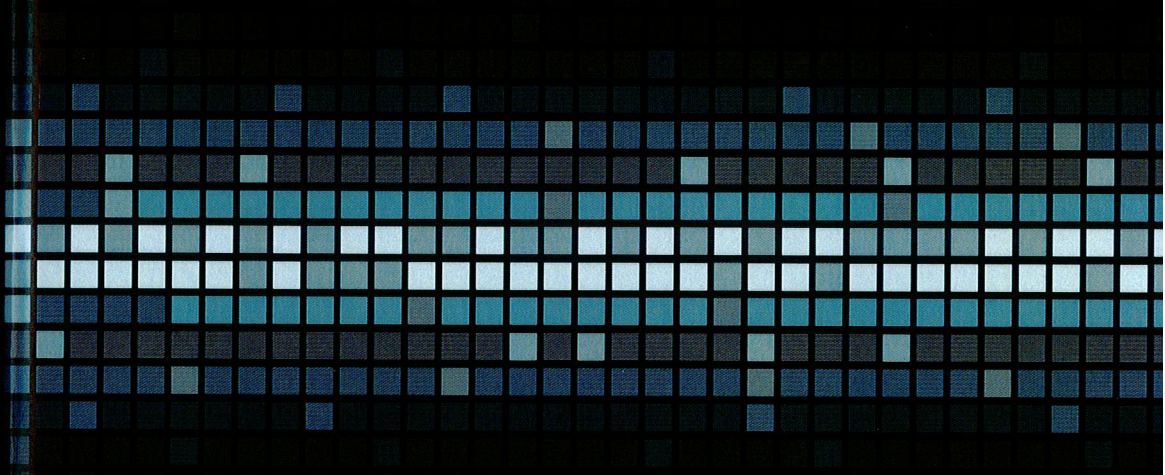
Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-49872
Book Section
Published Version

Originally published at:
Dürscheid, Christa; Stark, Elisabeth (2011). sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In: Thurlow, Crispin; Mroczek, Kristine. Digital Discourse. Language in the New Media. Oxford: Oxford University Press, 299-320.

OXFORD STUDIES IN SOCIOLINGUISTICS

DIGITAL DISCOURSE

Language in the New Media



EDITED BY

Crispin Thurlow and Kristine Mroczek

Digital Discourse

DIGITAL DISCOURSE

Language in the New Media

Edited by

CRISPIN THURLOW

KRISTINE MROCZEK

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc. publishes works that further
Oxford University's objective of excellence
in research, scholarship and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal
Singapore South Korea Switzerland Thailand Turkey Ukraine Vietnam

Copyright © 2011 by Oxford University Press, Inc.

Published by Oxford University Press Inc.
198 Madison Avenue, New York, New York 10016
www.oup.com

Oxford is a registered trade mark of Oxford University Press

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
electronic, mechanical, photocopying, recording, or otherwise,
without the prior permission of Oxford University Press,

Library of Congress Cataloging-in-Publication Data
Digital discourse: language in the new media / edited by Crispin Thurlow and
Kristine Mroczek.
p. cm. — (Oxford studies in sociolinguistics)
Includes bibliographical references and index.
ISBN 978-0-19-979543-7 (hardcover : alk.paper) — ISBN 978-0-19-979544-4 (pbk. : alk.paper)
1. Sociolinguistics. 2. Social media. 3. Digital media 4. Technological innovations—Social aspects.
5. Discourse analysis—Social aspects. I. Thurlow, Cripin. II. Mroczek, Kristine R.
P107.D54 2011
306.44—dc22 2010049086

1 3 5 7 9 8 6 4 2

Printed in China
on acid-free paper

For Sally Johnson

Contents

Foreword *xi*
Naomi S. Baron
List of Contributors *xvii*
Introduction: Fresh Perspectives on New Media Sociolinguistics *xix*
Crispin Thurlow and Kristine Mroczek

PART ONE: *Metadiscursive Framings of New Media Language*

1. Voicing "Sexy Text": Heteroglossia and Erasure in TV News Representations of Detroit's Text Message Scandal 3
Lauren Squires

2. When Friends Who Talk Together Stalk Together: Online Gossip as Metacommunication 26
Graham M. Jones, Bambi B. Schieffelin, and Rachel E. Smith

3. "Join Our Community of Translators": Language Ideologies and/in *Facebook* 48
Aoife Lenihan

PART TWO: *Creative Genres: Texting, Messaging, and Multimodality*

4. Beyond Genre: Closings and Relational Work in Text Messaging 67
Tereza Spilioti

5. Japanese *Keitai* Novels and Ideologies of Literacy 86
Yukiko Nishimura

6. Micro-Blogging and Status Updates on *Facebook*: Texts and Practices 110
Carmen K. M. Lee

PART THREE: *Style and Stylization: Identity Play and Semiotic Invention*

7. Multimodal Creativity and Identities of Expertise in the Digital Ecology of a *World of Warcraft* Guild 131
Lisa Newon

8. "Ride Hard, Live Forever": Translocal Identities in an Online Community of Extreme Sports Christians 154
Saija Peuronen

9. Performing Girlhood through Typographic Play in Hebrew Blogs 177
Carmel Vaisman

PART FOUR: *Stance: Ideological Position Taking and Social Categorization*

10. "Stuff White People Like": Stance, Class, Race, and Internet Commentary 199
Shana Walton and Alexandra Jaffe

11. Banal Globalization? Embodied Actions and Mediated Practices in Tourists' Online Photo Sharing 220
Crispin Thurlow and Adam Jaworski

12. Orienting to Arab Orientalisms: Language, Race, and Humor in a *YouTube* Video 251
Elaine Chun and Keith Walters

PART FIVE: *New Practices, Emerging Methodologies*

13. From Variation to Heteroglossia in the Study of Computer-Mediated Discourse 277
Jannis Androutsopoulos

14. *sms4science*: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland 299
Christa Dürscheid and Elisabeth Stark

15. C me Sk8: Discourse, Technology, and "Bodies without Organs" 321
Rodney H. Jones

Commentary 340
Susan S. Herring

Index 349

Chapter 14

sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland

Christa Dürscheid and Elisabeth Stark

Introduction

> FREED OF ANY restrictions imposed by grammar and spelling, cell phone users in Switzerland enjoy texting. Dialect is used alongside Standard German or French, words are omitted, shortened or creatively modified; English short-forms like *cu* (= see you) are being used, languages get intermingled. [...] Not all texters use these (and other) strategies, and those who do, abandon them depending on the situation. Our research group investigates which means of expression are actually used in SMS, which varieties of spelling are being used for one and the same concept (e.g. *bisous, bizous, bizoux, bx, b* for *bisous,* 'kisses' in French) and also the strategies used for typing fewer characters [...].

This statement is part of an announcement (translated into English) we published in Swiss newspapers, via broadcast and on advertising folders in September 2009 in order to invite Swiss people to send us their text messages.[1] The text makes allusions to features often assumed to be typical for text messaging. But how can we know whether these features are really applied, whether and how languages in Swiss text messages are intermingled, whether words are omitted and abbreviated forms are actually used? Maybe the use of certain writing strategies mentioned here (such as *cu*) depends on the age of the texters, maybe they do not appear at all?

Corpus-based research is capable of answering these questions, since it not only offers the possibility of checking hypotheses on writing strategies empirically but also allows for corpus-driven research, that is, finding patterns that have not yet been taken into consideration as special features of text messages within SMS ("short message service" or text messaging) communication. As we will show shortly, some corpus-based SMS studies already exist, but up until now the databases used were very small and the findings therefore not statistically significant.

This is why we launched the project *sms4science.ch*, a subproject of the international project *sms4science* (coordinated in Belgium), which brings together researchers from various countries in order to conduct corpus-based research on text messaging. Our approach will be presented in detail below; first, we will give a brief overview of different types of new media corpora and the classification criteria used to distinguish them. Then, we will introduce some projected research work based on our corpus, which will give answers to linguistic and sociolinguistic questions about the text messages and the texters. Finally, we will finish by briefly discussing the future of text messaging and of text messaging research.

Reviewing Corpus-Based SMS Research

In order to position our own corpus and explain what is innovative about it, we start by introducing a number of corpora used in the past together with their classification criteria. This enables us to identify the typical features of these corpora, such as the type of data included, the data's origin, and the availability/accessibility of the data.

As we mentioned above, our project was not the first to carry out corpus-based research on new media data. The website http://www.cmc-corpora.de, a supplement to an article by Beißwenger and Storrer (2008), provides a selected list of corpora (e-mail, chat, newsgroups, mailing lists) and offers a typology that we will follow for our classification. First, the authors differentiate between *project-related corpora* and *corpora for general use*. They note: "The former are compiled as an empirical basis for questions in a particular project, the latter do not directly pertain to a particular research project" (cf. Beißwenger and Storrer, 2008, p. 294). Second, they make a distinction between *corpora of raw data* and *annotated corpora*. The *Enron Email Dataset* (cf. http://www-2.cs.cmu.edu/~enron/), for instance, contains over half a million business-related e-mail messages but is not annotated, while the *Dortmund Chat Corpus* (cf. http://www.chatkorpus.uni-dortmund.de/), which comprises 511 protocols of chat communication, contains annotations for emoticons, nicknames, gender as given by the informant, and more. Parts of this corpus are freely available for research work on the internet while the main corpus is password protected and not for general use.

Applying these criteria further to the classification of SMS corpora, we observe that most SMS corpora discussed in the literature do not comprise data for open access. Exceptions are the *NUS SMS Corpus* collected by students at the University of Singapore,[2] which consists of about 10,000 messages, or a corpus collected for a study by Schlobinski et al. (2001), with "about 1,500 messages" sent by texters aged "younger than 12 years" up until "older than 30."[3] The latter was compiled for a specific study, but its form and accessibility make it available for other studies, too. The main Belgian *sms4science* corpus (cf. Fairon et al., 2006a, 2006b) and our own sister corpus are for general use as well. Unlike our own corpora, many existing SMS corpora are also project specific; that is, they were created for the restricted research goals of a particular researcher. Examples of this type of limited or restricted corpora include a 544-message corpus from Great Britain (Thurlow, 2003; Thurlow & Poff, 2011), one with 882 messages in Norway (Ling, 2005), and a South African corpus with 312 messages (Deumert & Masyniana, 2008). A relatively large project-specific corpus, consisting of 10,626 English messages from friends and family (aged 19 to 68), was compiled by Tagg (2009) for her doctoral research.

From a technical point of view, one more important distinction between existing corpora and our own corpus is that in all other SMS studies (except Fairon et al., 2006a, 2006b) text messages were transcribed by participants. In this regard, the *sms4science* corpus is quite unique: We asked participants to forward their text messages directly to a designated, free mobile number, thereby avoiding any transcription errors or deliberate modifications. Another important difference to most other corpora is the availability of our data. At the beginning of her thesis, Tagg (2009, pp. 10–18) presents a comprehensive overview of previous sociocultural studies of text messaging and describes the respective corpus specifications, which shows that most of the mentioned SMS corpora (including her own) are not accessible on the internet. This is most likely due to privacy issues. Data created on the internet such as online chat are freely available and can therefore be used and redistributed. SMS data on the other hand are normally not publicly available because they are stored on the users' cell phone only. If users agree to share their text messages

for scientific research, it is the researchers' responsibility to protect the texters' privacy, and the easiest way of achieving this aim is by not making the corpus public. Furthermore, the data are often of a very confidential nature and thus have to be dealt with in an extremely responsible way. Absolute confidentiality and anonymity must be assured, and the texters have to agree to having their messages, rendered anonymous, made available for publication. However, collecting text messages in cooperation with the texters also represents a major advantage over data publicly available on the internet such as chat protocols because once the texters are willing to cooperate, they will also very likely provide personal demographical information such as age, sex, or education.

Finally, we would like to mention one last distinctive linguistic feature of our corpus. Typically, data in SMS corpora are based on a single language depending on the country.[4] The *sms4science* corpus was intended to be multilingual from the outset. In Switzerland, there are four official languages (German, French, Italian, and Romansh) and—at least for the German, the Romansh, and partially the Italian speaking part of Switzerland—there are also different regional dialects, which are used in almost every situation of daily life (cf. Rash, 1998; Siebenhaar, 2006). Accordingly, any text messaging data collected in Switzerland are likely to show a considerable degree of language variation (official languages, Romance dialects, and Swiss-German[5]). A project such as ours therefore offers not only insight into the specific practices of a particular mode of communication (i.e., text messaging) but also an array of new information about the contemporary nature and status of language use in a multilingual country (compare this with, for example, Peuronen, Chapter 8, this volume).

To sum up, different types of existing corpora can be classified as in Table 14.1 (with an example for each type), which we base on the classification scheme in the study by Beißwenger and Storrer (2008), but supplemented with further criteria in order to characterize the corpora in more detail. As can be seen, none of the corpora presented here contain in-depth demographic data, even though for some the nicknames and the given age of the informants are known. As for the data source, some corpora offer data obtained directly from the informants, manually or automatically, and others consist of data taken from the internet. Concerning data availability, there is a difference between data accessible on the internet in a browser and data that must first be downloaded in order to be used. Having the data available for direct access on the internet not only guarantees it to be

Table 14.1: The corpus *sms4science.ch* in comparison with other corpora

		Schlobinski et al. (2001)		COSMA[6]	
Type of data	sms4science.ch **SMS (cell phone)**	SMS Corpus **SMS (cell phone)**	Enron e-mail dataset **E-mail**	e-mail corpus **E-mail**	Dortmund chat corpus **Chat**
For general use	X	X	X	-	X
Annotated	X	-	-	X	X
Multilingual	X	-	-	-	-
Anonymous	X	X	-	-	X
Laid out for linguistic purposes	X	-	-	-	X
Data source					
Obtained from users	X	X	-	X	
Acquired from the internet	-	-	-	-	X
Availability of the data					
Download	-	X	X	-	X
Browser	X		-	-	X
Password protected	X	-	-	-	(X)
Special features					
Search for "Regular Expressions"	X	-	-	-	X
Demographics	X	-	-	-	-

available to any type of computer with any type of operating system but also makes sure that it is presented in a form ready for research and not in a raw format. Thus, the researcher can focus on her project and does not have to struggle with setting up a computer-based working environment. On a similar note, retrieval tools such as "regular expressions"—a formal language based in computer programming allowing search for text patterns—are built right into the online research environment. This feature allows, for instance, the identification of different spellings (e.g., *hello, helo, hallo, hallooooo*), collocations, and more.

Building an International SMS Corpus Network: sms4science

In 2004, Cédrick Fairon and his research group at the Institute for Computational Linguistics CENTAL of the Catholic University of Louvain (UC Louvain, Belgium) launched a scheme called "Faites don de vos sms à la science" ("Donate your text messages to science!").[7] The motivation for this enterprise, described in more detail below, was that—in spite of the ever increasing public and linguistic interest in the topic—no corpora of a significant size and comprising authentic cell phone text messages were available at the time. The resulting first step toward establishing a large corpus of electronically and automatically gathered (i.e., not transcribed) text messages was initially restricted to the French-speaking part of Belgium (cf. Fairon et al., 2006a, 2006b) but enabled the group to develop a general methodology for message collection and establish protocols for the preparation of SMS corpora (e.g., anonymization, transcription, annotation). The corporate partners of the project, the most important national telecommunication companies, managed the technical aspects of data collection. The Belgium media spread the news, which enabled the project to reach a large part of the French-speaking population. After only two months, the results were impressive: more than 75,000 text messages gathered from about 3,200 persons (aged between 12 and 73, with 76% of them under 25), with 2,775 participants also having answered a biographical questionnaire.

This successful start encouraged the Belgian group to contact researchers interested in corpus linguistics, sociolinguistics, and language variation in order to build an international network and to establish more and mutually comparative corpora of text messages stemming from different countries and languages. Currently, this research network comprises fifteen universities in nine countries (Belgium, Canada/Québec, France, Great Britain, Greece, Italy, Romania, Spain, Switzerland), and an array of commercial partners. The network makes possible a wide range of sociolinguistic studies on topics such as variation in pluricentric languages (e.g., French in four different regions in France, La Réunion, Belgium, Canada, and in Switzerland) or multilingualism within countries (e.g., in Spain with universities from Catalonia and the Basque Country; in Switzerland with the University of Zurich from the German-speaking and the University of Neuchâtel from the French-speaking parts of the country).

Closer to Home: The sms4science.ch *Corpus*

In 2008, a collaborative agreement was signed between the Belgian group, the University of Zurich (Elisabeth Stark for French and Italian, Christa Dürscheid for German, later Matthias Grünert for Romansh) and the University of Neuchâtel (Marie-José Béguelin, later also Simona Pekarek-Döhler, both for French), marking the beginning of the subproject *sms4science.ch*. The Swiss researchers then contacted a corporate partner, *Swisscom*, to help with the technical part of the data collection. *Swisscom* provided a central mobile number connected to an automatic collection tool, where potential participants could send their original messages for free or for a small fee, depending on their service provider. The decision of whether or not to make a specific text message available would thereby always remain with the informants. Most people either forwarded their text messages or added the project's phone number as a second recipient of the original message. Participants thus sent individual messages that they selected themselves to the designated mobile number to then be included in the corpus. Finally, before launching the data collection itself (from November 2009 to January 2010 with one previous call in September 2009), the University of Zurich set up a quadrilingual website (www.sms4science.ch) and made contact with the media, resulting in a broad coverage across Switzerland.

At the time of writing, the results of our data collection have been promising. We have collected a total of 23,988 text messages sent by 2,627 different people (see Table 14.2). 18% of the messages originate in the French-speaking part of Switzerland (the *Romandie*). Most participants have had between 1 and 5 messages forwarded to the project, with 80 people sending more than 50 text messages each and one even sending 413 messages. About half of the participants also completed the

Table 14.2: Some raw facts about the corpus *sms4science.ch*

	Text messages	Words	Participants	Questionnaires	Sex	
					M	F
Total Number	23,987	About 480,000	2,627	1,308	477	831

Table 14.3: Age of the participants (of those who answered this question)

10–19	20–29	30–39	40–49	50–59	60–69	70+
245	599	190	152	80	38	5

Table 14.4: Mother tongues of the participants (more than one possible)

	Swiss-German	Standard German	French	Italian	Romansh	Others
Total number	889	161	256	54	26	125

biographical questionnaire, anonymously indicating their sex, age, profession, education, mother tongue, language competence, and specific SMS habits (e.g., frequency of use, language mixing, preferred addressees, use of other new communication forms like e-mail, online chat, etc.) as well as their general reading and writing habits (e.g., whether and how often they read and write in a week, what they read [newspapers, books, etc.], what they write, how they write [by hand, etc.]). Given this biographical information, we know that 831 participants are women (60%), that more than two thirds of those who filled in the questionnaire have Swiss-German as their mother tongue, 12% Standard German, 20% French, 4% Italian, 2% Romansh, and 10% other languages (more than one option was possible). These figures reflect quite accurately the percentages of the Swiss population in general. About 45% of our participants are between 21 and 30 years old, but we also have many teenagers and people over 50. (Seven people are older than 70.) Altogether, about 75% of our text messages can be linked to sociodemographic information (see Tables 14.3 and 14.4). This rate is higher than the 50% of the informants who submitted biographical details because of those people who sent in more than one text message.

With regard to the content and the form of the text messages, our initial impression is that they mostly comprise personal communication about romantic dates, problems, love affairs, and jokes, but we have also noticed messages of a more official character such as exchanges between business partners or between pupils and teachers (mostly apologies for absences). Even at first sight, their linguistic form and orthography shows several interesting features such as a high level of multilingualism and an overall preference for Swiss-German dialects as opposed to Standard German. It is even possible, from phonetic spelling, to recognize the different dialect regions of Switzerland. By *phonetic spelling* we mean that texters try to approximate the pronunciation of the word (cf. Frehner, 2008, p. 104; also Thurlow, 2003). Since the pronunciation (and many other linguistic features) may differ from dialect to dialect, the spelling may sometimes reveal the dialectal zone the texters are coming from, without having to consider the zip code in the sociodemographic information of the questionnaire.

Moving from Raw Data Toward a Linguistic Corpus

From the very beginning, our research group wanted the *sms4science* corpus to be open to other academics for different types of studies. This was not only a fundamental aim when collecting the data, but also when processing and presenting it. Unlike its Belgian counterpart, which is distributed as Microsoft Access® database on a CD-ROM, our Swiss corpus will be made available as an online database accessible with any major web browser (after having received a password form our research group, a security feature that allows us to restrict the access to researchers and students). In its current state, the corpus is already capable of processing "regular expressions" (see above), which will be greatly enhanced once the planned annotations (discussed further in this section) are added to the corpus.

The data, consisting of both text messages and biographical information, were collected by *Swisscom* and made available to the research project in the form of two independent mySQL dumps (SQL = *Structured Query Language*), the standardized export/import format of the open source database system MySQL.[8] Biographical data and text messages were linked by a hexadecimal code derived from the phone number but generated by SWISSCOM in order to keep the texters' privacy while still providing us with information about sex, age, and other demographics for each participant who filled in the biographical form. This cooperation with the

provider also ensured that text messages longer than 160 characters were kept together as one text. In the final version of the corpus, personal data are still stored in a mySQL database while the text messages are retrieved from an XML file.

Turning to the SMS database, four problematic aspects for the future usability of the corpus were immediately apparent. First, not all messages could be retained for the future corpus. There were a number of doublets (i.e., exactly repeated messages), which had to be removed, as well as some of the initial messages sent by participants in order to sign up for the project, all of which reduced the number of messages by about 7,000 to the quoted number of 23,988. All other marginal cases (i.e., automatic messages sent by computers such as information about incoming e-mail or else reminders for administrative actions, and so forth; messages not written but forwarded by the participants; messages written on a computer keyboard and not on a cell phone) were kept in order to give a complete and authentic picture of the data collection. The very first were exported into a separate list; the last ones, wherever recognizable, were marked in the corpus in order to identify them when working with the corpus, assuming that the different condition of production influences their linguistic and formal appearance. This was a slight deviation from the procedures of the Belgian group (cf. Fairon et al., 2006a, p. 19f.).

A second challenge we have faced is that messages contain a lot of confidential information such as names or phone numbers. This was first of all a legal problem and required that all confidential information in the text messages be eliminated systematically and, wherever possible, automatically (cf. for the Belgian corpus, Fairon et al., 2006a, p. 21).[9] So all types of numbers, that is, telephone numbers as well as street names and e-mail addresses were substituted. Numbers consisting of three or more digits were replaced by NNN, where every N stands for one digit. Also for e-mail addresses, the number of characters was kept while replacing them with xxx@yyy.ch. Street names were replaced as a whole by [StreetAddress]. We decided not to substitute toponyms, website addresses, and names of public institutions, economic enterprises, or brands since they are unlikely to reveal confidential information.

Personal names represented a special problem with regard to future research on our corpus, particularly in terms of sociolinguistic and communicative questions, such as who communicates how with whom. Around 95% of the (few) surnames in the corpus could be found manually and then automatically substituted by [LastName]. A different approach

was chosen for first names, thereby deviating from the approach of the Belgian team (cf. Fairon et al., 2006a, p. 21). Because dialogic sequences are easily retrievable through repeating first names, we decided to rotate the first names found within the corpus so as to detach names from content rather than to replace them by a label, for example [FirstName], so *Paul* would become *Ted, Fred* would become *Peter*, and so on. In cases of potential homonymy, items remained unsubstituted (e.g., the name *Hans* is homonymous with dialect quite frequently; e.g., *I hans gseh*, "I have seen it"). We consider the probability of a person being identified based on these very common first names to be close to zero. Additionally, because we did not replace first names but rotated them, the researcher working on the corpus will never know whether a first name he comes across has actually been rotated or not and will therefore not even attempt to match it to real persons.

As a result of all these anonymizations, we obtained messages such as the following one (in Standard German):

Extract 14.1:

Zur Erinnerung: diese Woche Mittwoch, 0.8:15 Berufsberatung [StreetAddress], Bottmingen, Tel: NNNNNNNNNN, Brief vorweisen. Und am Freitag, dem NNNNNNNNNN. um 12.50 zu Dr. [Lastname], Birshofklinik neben MFP Münchenstein. GlG! Mama

"As a reminder: this week Wednesday: 0.8:15 career counseling [StreetAddress], Bottmingen (= place name), Tel: NNNNNNNNNN, present letter. And Friday, NNNNNNNNNN. at 12.50 at Dr. [Lastname], Birshof-clinic, next to MFP (= public building), Münchenstein (= place name). GlG! (= very kind regards) Mama"

In this way, our corpus data were anonymized to a standard that complies with both Swiss legislation and the promises made to potential participants.

Third, our text messages contain different languages and language varieties, especially different forms of Swiss-German, with which we are not necessarily familiar and which have to be identified in order to permit a sufficiently accurate language tagging (e.g., only French text messages, only Swiss-German messages, etc.). Our intention was from the very beginning to preserve the corpus' multilingual character because multilingualism is an essential sociolinguistic fact of life in Switzerland. Given that a considerable number of text messages have already shown signs of code mixing, we

considered two possible forms of language tagging: (a) marking the whole SMS with the languages used in this SMS, and (b) marking the individual text parts in the respective languages. In Extract 14.2, we show the complexity of some multilingual text messages:

Extract 14.2:

<Spanish>**Olla fratello!!!**</Spanish> <Italian>**Come stai?**</Italian> <Standard German>**Wie geht's dir so? Immer noch so lange am arbeiten wie früher? Ich hab endlich mein eigenes Restaurant und**</Standard German> </Spanish>**mucho travajo**</Spanish> ...;-) </Standard German>**aber macht mir extrem spass**</Standard German> ...;-) <Italian>**allora amore, buona giornata**</Italian> <German Dialect>**und luegsch uf di, gäll**</German Dialect> ...;-) <English>**peace**</English>

"Hello brother!!! How are you? Still working as long hours as before? I finally have my own Restaurant and a lot of work ...;-) but it's extreme fun... ;-) so my love, have a nice day and take care, yeah...;-) peace"

We opted for the solution of attributing multiple language tags to individual messages, allowing users of the corpus to search for, for example, all messages containing French, including those that are multilingual, with French being one language among two or more. As a first step, a trained coder applied a tagging system for the main language in every SMS. For Standard German, French, Italian, and English, existing word lists were used to compare words in the corpus with words on the lists and thereby recognize the respective language. However, for Romansh and especially Swiss-German, no data were available. A former student of ours came to help. She had, in fact, created her own corpus of Swiss-German text messages for her Master's thesis and kindly made it available to us to be used as a word list. Additionally, our specialist for Romansh provided us with electronic newspapers containing texts in all the Romansh dialects and also texts in the standard variety *Rumantsch Grischun*.

This first automated tagging resulted in a language tag for each individual SMS, which is fairly accurate. In a next step, the aforementioned tagging with all languages contained in individual text messages will have to be tackled. For this step, classes will be offered to students of German and Romance philology and computational linguistics. In the course of these classes, students will be asked to tag all the messages for all the languages they contain; at the same time, they will also be asked to verify the

automatically applied main language. A special challenge in this procedure will be the assignment of loan words (e.g., French *merci*: "thank you" in, for example, Swiss-German). So the students will have to be guided by strict rules defined by the team.

Finally, many messages also show a considerable number of graphical variations. This remains an open problem for the moment and will be addressed in classes at the University of Zurich in the near future, too. The transcription of the original graphical representation (in at least four languages) in a standard orthographic form cannot be done automatically and comprises a multitude of decisions (cf. Fairon et al., 2006a, pp. 21–24 and pp. 100–110, for the Belgian corpus). There are several reasons for this. One is the very high level of graphical variation found in text messages. Compare the different graphic variants of *soirée* ("evening") or *bisous* ("kisses") in Extract 14.3:

Extract 14.3

a. Merci. Bisous, bonne soirée...
b. Bonne swarée et a+??? Bisouxx
c. Bizzøux bone soiré
d. G pase bon soire. Now g mal tet (Fairon et al., 2006a, p. 23)

Examples of the same phenomena, here based on the special writing strategy of using the phonic value of letters and numbers for homophonic syllables or words ("letter/ number homophones"), are very frequent in English (cf. the examples listed by Thurlow, 2003) and can be found in our corpus as well, for example, for *guet nacht* (Swiss-German "good night" using the phonic value in German of the number *acht* "eight") in *guetn8, guet n8, g n8,* or *gutN8*. In fact, only very few (French) items that deviate from standard spelling show regularities that point toward a form of typical "SMS spelling" (e.g., in the Belgian corpus this holds for <pcq> for *parce que*, "because," or <tt> for *tout*, "everything," never amounting to more than 70% of all occurrences, cf. Zimmermann, 2009, p. 130).

Another obstacle for an automated standard transcription (or coding system) is the fact that at least French text messages show some innovative structural phenomena like new conversions (denominal verbs without any markers). These are not easily understandable and recognizable by corpus users unfamiliar with specific linguistic phenomena of SMS and certainly not by the sorts of analytic tools used by computational linguistics. Take a look at the following extracts, for example (the first from Fairon et al., 2006a, p. 23):

Extract 14.4:
Mon prochain sms, li le qd tu **dodo** stp (= mon prochain sms, lis-le quand tu dors, s'il te plait)

"**my** next SMS, please read it when you are asleep/in bed"

Extract 14.5:
Je quitte la répét + tot.Ouf.Vais 1 peu mieux.Ai essayé de te **tel** mais pas de réponse.Bonne soirée cartes mon adorable amour. Moi **repas** puis **dodo** dès 22h. TQA (= Je quitte la répétition plus tôt. Ouf. Je vais un peu mieux. J'ai essayé de te téléphoner, mais il n'y a pas de réponse. Bonne soirée ??? mon adorable amour. Moi je vais manger/dîner et ensuite dormir à partir de 22 heures. TQA)

"I leave the rehearsal earlier. Ouf. I am a bit better. I have tried to call you, but there was no answer. Goodnight ??? my beloved one. I will have dinner and then go to sleep at about 22 o'clock. TQA"

In Extract 14.4, *dodo*, part of the phrasal verb *faire dodo*, "sleep," is used as a verb alone, which does not exist in this form in Standard or even colloquial French. Likewise, the abbreviation *tel* for (numéro de) *téléphone*, "telephone (number)," appears in Extract 14.5 as the abbreviation for *téléphoner*, "to call." And the elliptical *moi repas puis dodo* in Extract 14.5 has to be understood and/or transcribed as *moi je vais prendre un repas et ensuite faire dodo*, "I will have dinner and then go to sleep," thus omitting at least the verbs *prendre* and *faire* or converting the nominals *repas* and *dodo* into verbs. Finally, some quite regular omissions in French text messaging can be indicators of a certain informal style or code (like the omission of the first negation particle *ne* of standard French bipartite sentential negation *ne...pas*), while the omission of articles in front of nouns (similar to telegrams) are ungrammatical in every variety of French, just like in English. Accordingly, the former should be left out in a transcription, too, while the latter should be added. In the following example, *ne* in front of *oublie* ("forget") has to be left out also in the transcribed version of the SMS, as it would never appear there in spoken French:

Extract 14.6:
1h juste 1h après tu rentre s'il te plais et oublie **pas** le carton (= Une heure, juste une heure, et après tu rentres, s'il te plaît, et n'oublie pas le carton)
"one hour, just one hour and then you'll come home, please and don't forget the carton"

What becomes clear from all this is that any standardized transcription of original text messages must inevitably be subject to some interpretation by the transcriber, and clear reasons must be given for choosing one or the other variant.

The last step in the constitution of our corpus is the implementation of *Corpus Navigator*, a corpus-browsing tool developed at the English Department of the University of Zurich.

In addition to the main corpus, some smaller corpora will be available through the same website. One will be the corpus of machine-generated messages (as previously mentioned), and another one a private collection of some 1,200 dialogical messages we received. These messages cannot be included in the main corpus, first due to their different production/transmission conditions, and second because of a potential overrepresentation of certain ideolectal features, given the high number of single text messages written by only two individuals compared with the rest of our corpus. Yet, they may be considered ideal for studies on a dialogical level, so we still want to make them available.

In the final version of *Corpus Navigator*, the software should allow for searches for single items, including emoticons, different strings of graphic characters, sociodemographic properties of the respective texters, and possibly also parts of speech (which of course presupposes as an additional step a part-of-speech-tagging; cf. for the Belgian corpus, Fairon et al., 2006a, pp. 25–30).

Taking sms4science *Forward: Looking to the Future*

In the next few years, a series of different graduate student research projects (e.g., master's and Ph.D. theses) will be conducted through third-party funded research on the basis of our corpus. In bringing this chapter to a close, therefore we want to highlight a few of the kinds of possible research questions we envisage. These concern (a) language choice and code switching; (b) structural features of text messages; and (c) pragmatic issues. Thanks to the demographic data submitted by the participants in our study, we have a solid empirical foundation for investigating these types of topics.

(a) Investigating Language Choice and Code Switching

Given the fact that Switzerland is a quadrilingual country, there are at least three main research fields concerning the messages at hand. The first of these is the degree of code switching different to that in monolingual

countries, and this question: What are the communicative reasons for code switching when it does take place?[10] One respective hypothesis might be that code switching is particularly frequent in the Swiss corpus and much less so in officially monolingual countries. On the other hand, Swiss speakers themselves are mostly not multilingual, their main language usually is the language of the region/the canton in which they live. It will therefore be worthwhile to compare our data with the text messages gathered in other countries participating in the *sms4science* project to investigate the influence of a multilingual environment on the individual monolingual texter. The second research field we can identify is this question: Which varieties of the four national languages in Switzerland are actually used in the Swiss SMS corpus? Is it true, for instance, that speakers of Swiss-German mainly draft their messages in dialect and not in Standard German (cf. the investigation of Braun, 2006)? By the same token, one might ask if there is a correlation between the age of texters and their choice of Standard German or Swiss-German. As Siebenhaar (2006, p. 492) has shown for Swiss internet chat rooms, younger chatters use more dialect, while the middle-age generation prefers Standard German. Does this apply for SMS communication in the German-speaking part of Switzerland as well? Lastly, our third research field concerns the following: And what about messages submitted in Romansh? Are they usually monolingual, or do we always find two languages within these messages given the fact that there are no monolingual Romansh speakers? One may, for instance, assume that typical abbreviations of a second language (such as *hdl* in German, "hab dich lieb," "I love you") are used in the text messages even if the dominant language is Romansh.

(b) Investigating Structural Features of Text Messages

On a structural level, future research projects will examine the persistent issue of orthography, for example, the use of abbreviations, spelling practices in general, strategies of phonetic writing (searching for occurrences of *cu* for *see you*, *kul* for *cool*, emoticons, etc.), letter–number homophones, nonstandard spelling, and the use of uppercase lettering (such as *SUUUPER*). Furthermore, two kinds of ellipses will be analyzed, especially in relation to the morphosyntactic and dialogical structure of the messages at hand. In the first type of ellipses, functional elements (such as articles) are omitted; in the second one, content words are dropped. The latter typically occurs in responses to previous messages.[11] Both are features commonly assumed to be typical for text messaging and ones that can be empirically

Your query returned 7510 results in sms.

|< << >> >| Show Page: 3 Sentence View New Query Go!

No	Sender ID	SMS ID	Solutions 41 to 60 Page 3/376 Processed for		
41	14	94	ganz knapp am epa-platz verwütscht!	;-)	dicke kuß!
42	14	95	denn langsam mal em waßer nöchere!	;-)	riesequalle -wo's luut 20 min da sött ha- hend mer uf jede
43	14	95	hend mer uf jede fall no keini gseh!	;-)	ich hoff, du hesch no alli 10 finger und chasch hüt au no c
44	2007	97	würd bstelle! Isch das na möglich	;)	? Glg und n schöne Sunntig na! Viola
45	15	99	hüt abig ah, wenn ih es nid vergisse.	;)	Danke, dass du um mi sorge gmacht hesch und für mi do bisch
46	15	99	sehr ah dir. Hab dich ganz doll lieb.	=)	*knuddel* Lg Reinhard (is back in town)
47	2012	117	bis 13 uhr und darf nicht tel. Sms geht	;)	Lg gerhard
48	17	120	Heii schätzli	:-)	Ja ha e guete start gha, abr ha schowidr huere dr aschiß. U
49	17	120	Ha di letscht wuche nid einisch gseh	:-(	vermiße di! KiZz <3
50	19	126	doch en vorschlag, mir isches gliich	:-)	
51	21	131	i miss you! Du hesches bald gschafft	:-)	vernünftig bisch mitem zug gange, super! Cu soon :-*
52	21	131	bisch mitem zug gange, super! Cu soon	:-*	
53	2018	133	chume sicher. Tönt guet mittem Chnobli	:-)	gruß Niels
54	2019	136	oder zersch chile und den shope?	:)	Wer voll sozi vo dir x) Mfg eugen
55	2019	138	vo hüt erhole, ha viel"dezueglernt"	;)	Lg & gn8 Ps: ma luege denke aber ehner weniger.. We'll see!
56	2019	142	Meci	:)	stahne scho am bahnhof xD Gibder es phone wani ufde zug gah
57	2022	146	gleid und wünsche dir en guete tag.	:-)	
58	2022	149	Danke esch lieb vo dir	:-*	kuß
59	24	151	meine Maus Tot ist! und das stimmt!!!	;-(	
60	2019	155	a wend am abelaufe bisch Mfg eugen	=)	

Corpus Navigator 2.5 © 1989–2008 H.M.Lehmann

FIGURE 14.1 Emoticons in a KWIC view representation in Corpus Navigator.

investigated in our SMS corpus, but also framed theoretically. Due to the retrieval tools in *Corpus Navigator*, it is also possible, for instance, to find all the orthographic variations of one word and to gain an idea about the frequency of specific items.

Here is an example of what the results of a search for emoticons in *Corpus Navigator* might look like (Figure 14.1). The illustrations show in which context emoticons appear and how often they can be found all together (i.e., 7,510 times).

By examining these types of features, it will be possible to verify popular stereotypes (cf. Thurlow, 2006, for an overview) about texting in a well-grounded, empirical manner. It will also be possible to reveal underlying regularities in the messages, showing that the morphosyntax of texting does not diverge from language-specific or maybe even universal rules (cf. Stark, 2011). Finally, thanks to the sociodemographic data, it will be possible to examine correlations between stylistic choices and age, sex, or the educational background of the texters.

(c) Investigating Pragmatic Issues

As we have already mentioned, within *sms4science.ch* some dialogically oriented corpora will be available as well. These subcorpora can serve as

a basis for the study of text messaging conversations, the study of the relationship between the interlocutors, or the study of the similarities and differences between texting and face-to-face communications. However, this type of discourse-analytic research may not only be examined in the subcorpora. Some of the questions can also be answered by the main corpus because the texters can partially be recognized by their identifiers and first names as being one and the same person, thus making it possible to retrace the respective messages.

Another pragmatic research field is the use of salutation formulas at the beginning and the end of the messages (see Spilioti, Chapter 4, this volume). Is it correct that they are often missing, in other words, that "the most significant characteristic of salutations is their absence" (cf. Frehner, 2008, p. 91)? And are there any differences between individual languages in using or omitting these formulas? Additionally, the main communicative functions in the text messages may be investigated. Regarding this research field, several possible questions spring to mind: Does the thematic content (personal matters, business) correspond with the degree of informal or formal style used in the messages? Are the particular strategies applied to maintain a social relationship (such as sending good night messages) the same in all languages in our corpus, and are they the same in the international SMS corpus, which will be established on the basis of the individual national ones?

The Future of SMS—A Guessing Game?

No one can predict whether text messaging will remain as frequently used as it is at the moment, especially given the fact that there are other communication tools that can be used on cell phones as well (such as social networking sites, instant messaging, and microblogging such as *Twitter*). This question clearly has implications for future research. What, for example, are the differences between text messaging and microblogging? Are they substantially different communication practices, or is this just old wine (i.e., text messaging) in new bottles (i.e. *Twitter*), since in both cases we are faced with messages limited to a certain number of characters and since one-to-one messages are possible within *Twitter* as well? (See Lee, Chapter 6, this volume.) If the internet can be accessed from a mobile device, *Twitter*, *Facebook*, instant messaging, and others may be used on the cell phone as well, so the differences seem to minimize even more. And like *Twitter* or other modes of communication, text messaging is not necessarily text based, as other modalities (such as images and video clips) may appear in text messages as well.[12]

As for the frequency of use of texting compared with other modes of communication, a *Pew Internet & American Life* study from April 2010 revealed that the use of text messaging had continuously been gaining ground in the USA with one in three teens sending more than 100 text messages a day (Lenhart et al., 2010).[13] Given this development in one of the richest countries on the planet, we can assume that there will remain a huge interest in all kind of research work around SMS communication. However, it is imperative that the scientific exchange between scholars in different countries be intensified in order to carry out better-informed research on text messaging. As we have already seen, different SMS corpora exist, which were compiled for different purposes. Furthermore, studies in SMS communication other than in English should be brought closer to the research community. Perhaps *sms4science* will achieve the goal of bringing together SMS researchers from all over the world.

Acknowledgments

We would like to thank the research group of Cédrick Fairon, especially Louise-Amélie Cougnon, for their continuous support of the Swiss subproject; our Swiss colleagues from the University of Neuchâtel, Marie-José Béguelin and Simona Pekarek-Döhler, for their always pleasant and most effective cooperation; *Swisscom*, and especially Peter Schüpbach, for the perfect technical support of our data collection; and, finally, all our collaborators at the University of Zurich, particularly Charlotte Meisner, Andi Gredig, Beni Ruef, and Hans Martin Lehmann. We also thank our two reviewers and Crispin Thurlow for their very helpful comments. A special thanks goes to Simone Ueberwasser for her fruitful ideas, technical support, proofreading, and correction, and for always keeping a clear view whenever we were lost in technical or administrative trouble. All remaining errors and shortcomings are, of course, ours.

Notes

1. Note that in Switzerland, according to *Swisscom* (one of the major Swiss telecommunication companies), on their network alone 10 million text messages are being sent per day. This is an impressive number for such a small country with around seven million inhabitants only, and the use of SMS is still on the rise.

2. Cf. http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/
3. The demographic information on the texters as well as the total number of text messages is only given in a very general way (cf. http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf).
4. Cf. Fairon et al. (2006a, p. 100), who decided to erase all non-French data from their Belgian corpus, even though Belgium is a bilingual country.
5. Swiss-German is the "umbrella term applied to all German dialects spoken in Switzerland" (Rash, 1998, p. 21).
6. COSMA (= Cooperative Schedule Management Agent) is a German e-mail corpus containing 160 messages, cf. ftp://lt-ftp.dfki.uni-sb.de/pub/papers/local/klein97_dgfs.ps.gz
7. Cf. the website of the project at the time of data collection: http://www.sms-pourlascience.be.
8. This format was considered to be superior to the most obvious format for such a task, a text file in which the fields are separated by one or several defined characters, because basically any combination of characters must be expected in a text message, and therefore no combination of any characters at all can be used as a separator.
9. For ethical considerations and the principles of anonymization, cf. also Tagg (2009, pp. 80–93).
10. Siebenhaar (2006) discusses code switching between Standard German and Swiss-German, that is, between two varieties of the same language. For a broader view on code switching, we refer the reader to Chapter IV in the volume "Multilingual Internet" (2007), edited by Brenda Danet and Susan C. Herring.
11. For this type of adjacency ellipsis, cf. Klein (1993).
12. Although this is not the case in our SMS corpus.
13. When the study was published, a Swiss newspaper referring to this data was titled: "US-Teenager sind süchtig nach SMS" ("US-Teens are addicted to SMS", cf. Tages-Anzeiger, April 22, 2010).

References

Beißwenger, M., & Storrer, A. (2008). Corpora of computer-mediated communication. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, vol. 1, (pp. 292–308). Berlin & New York: de Gruyter.

Braun, B. (2006). Jugendliche Identitäten in SMS-Texten. In C. Dürscheid & J. Spitzmüller (Eds.), *Zwischentöne. Zur Sprache der Jugend in der Deutschschweiz*, (pp. 101–114). Zürich: Verlag Neue Zürcher Zeitung.

Danet, B., & Herring, S.C. (Eds.). (2007). *The Multilingual Internet: Language, Culture, and Communication Online*. New York: Oxford University Press.

Deumert, A., & Masinyana, S.O. (2008). Mobile language choices – The use of English and isiXhosa in text messages (SMS). *English World-Wide*, 29(2), 117–147. Amsterdam: Benjamin.

Fairon, C., Klein, J.R., & Paumier, S. (2006a). *Le langage SMS. Etude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science.'* Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Fairon, C., Klein, J.R., & Paumier, S. (2006b). *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation*. CD-ROM. Louvain-la-Neuve, Belgium: Presses universitaires.

Frehner, C. (2008). *E-mail – SMS – MMS. The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern: Peter Lang.

Klein, W. (1993): Ellipse. In J. Jacobs, A. von Stechow, W. Sternefeld, & T. Vennemann (Eds.), *Syntax. Ein internationales Handbuch zeitgenössischer Forschung /An International Handbook of Contemporary Research*, (pp. 763–799). Berlin: de Gruyter.

Lenhart, A., Ling, R., Campbell, S., & Purcell, K. (2010). *Teens and Mobile Phones*. Pew Research Center's Internet & American Life Project, April 20, 2010. Retrieved July 13, 2010, from http://www.pewinternet.org/~/media//Files/Reports/2010/PIP-Teens-and-Mobile-2010.pdf

Ling, R. (2005): The socio-linguistics of SMS: An analysis of SMS use by a random sample of Norwegians. In R. Ling, & P. E. Pedersen (Eds.), *Mobile Communications: Re-negotiation of the Social Sphere* (pp. 335–349). London: Springer.

Rash, F.J. (1998). *The German Language in Switzerland: Multilingualism, Diglossia and Variation*. Bern: Peter Lang.

Schlobinski, P., et al. (2001). Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen. Aspekten in der SMS-Kommunikation. *Networx*, 22. Retrieved July 13, 2010 from http://www.mediensprache.net/de/networx/docs/networx-22.asp

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics*, 10(4), 481–506.

Stark, E. (2011). La morphosyntaxe des SMS suisses français: Le marquage de l'accord sujet-verbe conjugué. *Linguistic Online*.

Tagg, C. (2009). *A Corpus Linguistics Study of SMS Text Messaging*. Unpublished Ph.D. dissertation at the Department of English, University of Birmingham, UK. Retrieved July 13, 2010, from http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text messaging. *Discourse Analysis Online*, 1(1). Retrieved July 13, 2010 from http://faculty.washington.edu/thurlow/papers/Thurlow%282003%29-DAOL.pdf

Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer Mediated Communication*, 11(3) article 1. Retrieved July 13, 2010 from http://jcmc.indiana.edu/vol11/issue3/thurlow.html

Thurlow, C., & Poff, M. (2011). Text messaging. In S.C. Herring, D. Stein, & T. Virtanen (Eds.), *Handbook of the Pragmatics of CMC*. Berlin & New York: Mouton de Gruyter. Retrieved July 13, 2010, from http://faculty.washington.edu/thurlow/papers/thurlow&poff%282009%29.pdf

Zimmermann, T. (2009). Le 'langage SMS' – une nouvelle varieté écrite de la langue française? Une analyse empirique basée sur un corpus de 30'000 SMS sous considération particulière de la relation phonie-graphie. Unpublished master's thesis, directed by Elisabeth Stark, University of Zurich.